

DOCUMENT RESUME

ED 338 630

TM 013 083

AUTHOR Young, John W.
TITLE Adjusting the Cumulative GPA Using Item Response Theory.
PUB DATE 89
NOTE 20p.; Doctoral Dissertation, Stanford University.
PUB TYPE Dissertations/Theses - Doctoral Dissertations (041)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Admission Criteria; College Entrance Examinations; *College Freshmen; *Equated Scores; Factor Analysis; Grade Point Average; *Grading; Higher Education; *Item Response Theory; *Predictive Validity; Universities
IDENTIFIERS Graded Response Model; Partial Credit Model; Stanford University CA; *Universal Scale for Grades

ABSTRACT

The development of a valid measure of academic performance through the application of existing statistical methodology in a new manner was studied. Focus was on analyzing the predictive validity of traditional pre-admissions measures in determining overall academic performance in college. This new criterion, termed a Universal Scale for Grades, is used to determine the predictive validity of certain pre-admissions measures for a particular cohort of students at one university. Two item response theory (IRT) models--G. N. Masters' (1982) Partial Credit Model, and F. Samejima's (1969) Graded Response Model--were used in this study. As far as known, this study represents the first application of IRT methodology to the problem of equating grades from different courses to a common scale. Data were taken from the undergraduate admissions office and the registrar's office at Stanford University. The sample consisted of all 1,564 students entering Stanford as freshmen in the autumn quarter of 1982. The database consisted of verbal and mathematics scores on the Scholastic Aptitude Test, College Board Achievement Test scores, high school grade point averages, rank in and size of high school class, and admissions staff assigned academic rating and non-academic rating. Only those records in course files that contained a letter grade were used. Principal factor analysis and validation results indicate that use of the IRT models to create more reliable measures of academic performance shows promise. (TJH)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

✓ This document has been reproduced as
received from the person or organization
originating it.

□ Minor changes have been made to improve
reproduction quality.

• Points of view or opinions stated in this docu-
ment do not necessarily represent official
OERI position or policy.

ADJUSTING THE CUMULATIVE GPA USING ITEM RESPONSE THEORY

JOHN W. YOUNG
GRADUATE SCHOOL OF EDUCATION
RUTGERS UNIVERSITY

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

JOHN W. YOUNG

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

INTRODUCTION

Criterion-related validity is a central concept in educational and psychological measurement. We are often interested in the relationship of a given predictor or set of predictors to some outcome measure. That is, how valid or how useful are these predictors in estimating some criterion? In institutional research on undergraduate admissions, the predictive validity of preadmissions measures such as standardized test scores and high school grades in determining academic performance in college is of interest to students, admissions officers, university administrators, educational researchers, guidance counselors, and the general public.

The criterion most frequently used in studies designed to assess the predictive validity of measures used in college admissions has been the freshman-year grade point average (GPA) (Wilson, 1983). Literally thousands of validity studies using freshman year college grades as the outcome of interest have been conducted. Fewer studies have been done that examine the validity of preadmissions measures in predicting other criteria of success.

For a number of reasons, the first-year GPA is the criterion of choice to the neglect of almost all other academic outcomes. The first-year GPA is favored in institutional studies because it is a well-defined criterion, it is easily obtained from university records, and it is available relatively soon after the matriculation of a class of students.

Numerous validity studies using freshman year college grades as the outcome of interest have been conducted. Results of these validity studies have been summarized periodically over the past half-century (e.g., Fishman and Pasanella, 1960; Schrader, 1971; Ramist, 1985). The most commonly used predictors of freshman year GPA are admissions test scores (usually SAT-V and SAT-M scores) and high school GPA or high school rank in class.

Wilson (1983) located few studies involving prediction of cumulative GPA for his review. Clearly, the first-year GPA is neither a sufficient nor adequate sample of a student's level of academic achievement. Many

This study is the doctoral dissertation of the author which was completed at the School of Education at Stanford University. He is now an Assistant Professor of Educational Statistics and Measurement at Rutgers University, New Brunswick, New Jersey 08903-5050.

questions remained unanswered concerning the validity of preadmissions measures in predicting longer-term academic performance in college.

On the surface, a criterion such as the student's four-year cumulative GPA would appear to offer several advantages over the widely used criterion of first-year GPA. The four-year cumulative GPA has greater face validity and greater content validity as a measure of overall achievement than first-year grades alone. In addition, the cumulative GPA should exhibit better reliability than the first-year GPA since it represents a larger sample of each student's behavior.

Few studies involving prediction of cumulative GPA have been conducted because the four-year GPA is known to be a problematic criterion. Although the cumulative GPA maintains its reliability over time, its construct validity is open to question. That is, does the cumulative GPA maintain the same essential meaning and comparability from student to student? The statistical properties of the cumulative GPA would not be adversely affected if it is the case that either students were randomly assigned to courses or that students chose their courses at random. However, we know that neither of these situations occur.

The construct validity of a grade scale is diminished if the grades of some students tend to be spuriously increased for reasons unrelated to their academic performance, while grades of other students were spuriously decreased. Because grading standards vary from instructor to instructor within a department and from department to department within a university, distortion of the grade scale may occur. With instructors the potential distortion is due to arbitrary grading habits. The same is often true of departments since some departments, on average, award higher grades than other departments.

It follows that self-selection by students into particular courses influence the grades ultimately received by any given individual. Each course requires a unique combination of knowledge and skills. Because the choice of courses varies considerably from student to student, the cumulative GPA measures different productions of knowledge and skills for different students. Thus, the cumulative GPA does not consistently measure the same set of attributes for each student.

By self-selecting into courses or majors where high grades are easier to obtain, students can spuriously raise their cumulative GPA. Conversely, students who take most of their courses in departments with stricter grading standards will, in general, have lower cumulative GPA's.

Previous attempts to adjust grades for differences in departmental and school standards have, in general, not been successful (see e.g., Linn, 1966; Boldt, 1976). These prior grade adjustments have been based on differences among registrants in scores on a common examination. However, only linear transformations of predictor and/or criterion scores were attempted. The principal references are the work of Bloom and Peters (1961) and Tucker's (1963) central prediction system, which utilizes a series of multiple regression equations to obtain adjustments of both

predictor and criterion scores.

At the present time the predictive validity of standardized tests such as the SAT is open to debate. However, prior studies have been limited by the use of the freshman-year GPA as the criterion by which all predictors are measured. For lack of a better longer-term indicator of academic performance, the question of the predictive effectiveness of preadmissions measures remains unanswered.

The focus of this study is the development of a valid measure of academic performance through the application of existing statistical methodology in a new manner. This new criterion, termed a Universal Scale for Grades, is used to determine the predictive validity of certain preadmissions measures for a particular cohort of students at one university.

Development of a new criterion of overall academic success, one which is resistant to the effects of differential grading among instructors and departments, enables us to more accurately estimate the true validity of preadmissions measures in predicting collegiate academic performance. The term 'true' is used in the statistical sense of unbiased estimation.

The technology for developing a Universal Scale for Grades derives from item response theory (IRT). Two IRT models are used in this study. The first model is Masters's (1982) Partial Credit Model (PCM). The second is Samejima's (1969) Graded Response Model (GRM).

As far as known, this study represents the first application of IRT methodology to the problem of equating grades from different courses to a common scale. The PCM and the GRM are appropriate models for these data because it appears likely that their underlying assumptions can be met.

Data for this study originated from the undergraduate admissions office and the registrar's office at Stanford University. The sample under investigation consists of all students who entered Stanford as freshmen in Autumn Quarter 1982, nominally the Class of 1986. The sample size is 1564 students and includes those individuals who left the university before obtaining a degree.

The data base from the undergraduate admissions office consists of the following information for each student: SAT-V and SAT-M scores, College Board achievement test scores, HS GPA, HS rank in class, HS class size, and admissions staff assigned academic rating and non-academic rating.

The data base from the registrar's office contains the following information for each student: sex, ethnicity, declared major(s), degree(s) received, and data on all courses taken while enrolled at Stanford. The data on courses consist of these four components for each course: the quarter and year the course was taken, the department and course number, the grade received, and the number of units awarded. The two data bases were merged using a scrambled student ID number, and all individual

identifying information was removed.

THE STATISTICAL MODELS

This study is a new application of IRT to data generated from a source other than standardized tests. The data for this study correspond to an IRT framework in the following manner. Each student can be considered to be the equivalent of an examinee. Each Stanford course can be considered to be equivalent to a test item with polytomous responses represented by the grade earned in that course. Since students, primarily because they major in different fields, enroll for different combinations of courses, we are faced with a situation analogous to that found in many testing programs where examinees attempt different sets of items. IRT is especially well-suited for handling this problem of creating a common metric regardless of which courses or items were taken.

The use of IRT in this study differs from the use of IRT in standardized testing in a significant way. In most testing situations different sets of items are given to randomly equivalent groups of examinees. This randomization procedure allows for causal inferences to be drawn from the results. Random assignment of items to examinees meets Rosenbaum and Rubin's (1983) definition of strongly ignorable treatment assignment.

In this study, students self-select into courses of their own choosing. Random assignment of students to courses is not a possible condition. Instead, we have an observational study since there is no randomization mechanism for assigning students. Thus, we do not know what impact any possible student by course interactions may have on the results. The effects of this important difference between the two situations is not known but this distinction should be kept in mind.

It is not reasonable to assume that the grades from all courses in an institution as diverse as Stanford would be determined by a single ability trait or construct. If it is the case that performance in some courses are closely related to one ability factor and performance in other courses are related to other ability factors, then each subset of courses should be treated independently.

Principal factor analysis is used to partition courses into more nearly unidimensional subsets, each of which is scaled separately. This partitioning of courses allows the use of unidimensional models such as the PCM and the GRM. Unidimensionality is a restrictive assumption, one that is rarely met by real data. Instead, data are presumed to be approximately unidimensional if the data are influenced by a single dominant trait or factor.

In addition, both the PCM and the GRM assume that the data can be scored in two or more ordered categories. If the data come from tests, then it is presumed that each item has a response format that produces several ordered levels of performance. The number of levels is free to

vary from item to item. Note that the categories or levels are assumed only to be ordered; there is no restriction that these categories must conform to an interval scale. Neither the PCM nor the GRM is restricted to use with data from test items. The assignment of grades in courses are a natural extension of the idea of ordered levels of performance.

The PCM takes its name from the fact that some responses to items can be considered 'partially correct' and therefore should be given partial credit. The one-parameter Rasch model can be considered to be a special case of the PCM where there are just two levels of performance. The PCM was developed by Masters (1982) and is a latent trait model for data scored in two or more ordered categories, and shares the important property of parameter separability with other Rasch models.

The PCM has the following form:

$$\pi_{nxi} = \frac{\exp \sum_{j=0}^k (\beta_n - \delta_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^k (\beta_n - \delta_{ij})}, \quad x=0,1,\dots,m_i$$

This expression is the probability of person n scoring x on item i . This probability is determined by the person's ability parameter and the difficulties of the m_i steps in item i . β_n represents the ability of person n and δ_{ij} is the difficulty of the j -th step for item i . For an item with $m+1$ ordered categories, it is convenient to think in terms of the ' m ' steps which need to be taken in order to successfully complete the item (Masters, 1982).

The PCM is closely related to Samejima's (1969) Graded Response Model. The PCM and the GRM have the following important advantage over models for dichotomous scoring: When data can be scored in more than two ordered categories, this leads to a more precise estimate of an individual's ability than can be obtained by scoring data dichotomously. There are, however, some important differences between the two models.

The GRM is also conceptualized for data that occur in three or more ordered categories. In addition, the principle of local independence is assumed to hold. The operating assumption of the GRM is that when an individual encounters a test item, a latent variable is induced. The probability that this variable takes on a value greater than the k -th 'category boundary' depends on the person's ability, the value of the k -th item boundary, and the item's discrimination.

In most applications of the GRM, it has been common to assume that discrimination parameters are equal across all response categories. This is called the homogeneous case of the GRM.

The Graded Response Model has the following form:

$$\pi_{nki} = \frac{\exp (\alpha_i (\beta_n - \lambda_{ik}))}{1 + \exp (\alpha_i (\beta_n - \lambda_{ik}))},$$

This expression is the probability of person n scoring k on item i , where α_i is the discrimination parameter of item i , β_n is the ability parameter of person n , and λ_{ik} is the value of the k -th item boundary between categories k and $k+1$ associated with item i .

An important difference between the PCM and the GRM is in the meaning of the item parameters in each model. In the PCM, the j -th item 'step' δ_{ij} is the point on the ability scale where the probability curves for categories $j-1$ and j intersect for item i . In the GRM, the k -th item 'boundary' λ_{ik} is the point on the ability scale where responding in or above category k for item i is 0.5.

The Graded Response Model does not provide a simple general expression for the probability π_{nki} of person n responding in category k to item i . Instead, this probability is obtained by subtracting cumulative probabilities for all of the categories. In addition, the form of the Graded Response Model does not allow for the algebraic separation of person and item parameters. Thus, no sufficient statistic can be derived for either the person or item parameters in this model.

The PCM has an alternative conceptualization which enables sufficient statistics to be computed for both examinees and items. The item parameters in the PCM represent the difficulty of each successive step in the item. Instead of calculating the probability of an examinee responding in a particular category as in the Graded Response Model, the PCM is a statement about the probability of achieving category $k+1$ given that the examinee has reached category k . Using this approach, the PCM satisfies the requirements of a Rasch model. See Masters (1982) for a full mathematical development of the PCM.

MSTEPS and ORDINAL are FORTRAN computer programs which are used for parameter estimation. MSTEPS implements the PCM while ORDINAL, written by Muraki (1988), is used in conjunction with the GRM. MSTEPS is a commercially available program while ORDINAL has not yet been released for public use. A detailed description of the estimation methods used in MSTEPS as applied to a Partial Credit analysis is given in Chapter 4 of Wright and Masters (1982). The rationale in using both MSTEPS and ORDINAL is to compare the ability and course parameter estimates resulting from fitting two different ordered response models.

METHODOLOGY

In this study the simultaneously scaling of both letter grades and non-letter grades (i.e., grades of Incomplete, No Credit, Pass, etc.) would require some arbitrary and questionable mapping of these two types of grades onto a common scale. Thus, only those records in the course file which contain a letter grade (i.e., the grade scale ranging from A+ to D-) were used in this study. A dichotomous Pass/No Credit scaling would obviously be inadequate for purposes of this study. Note that the traditional GPA is also based upon letter grades exclusively.

For the purposes of IRT parameter estimation, letter grades are collapsed into fewer categories. For a given course, collapsing grades into fewer categories translates into fewer parameters to estimate. Although the estimates are less precise, because we are taking a more finely divided scale and collapsing to a more crude scale, the benefits derived in simplifying the problem are substantial.

The collapsing of letter grades was accomplished in the following manner: A+, A, and A- are collapsed into the A category; B+, B, and B- are collapsed into the B category; C+, C, and C- are collapsed into the C category; and D+, D, and D- are collapsed into the D category. The letter grade scale is converted to a numerical scale using the Stanford registrar's system whereby A=4, B=3, C=2, D=1. In the registrar's system, plus/minus grades modify the letter grade by $\pm .3$.

Initially, data analysis and parameter estimation were restricted to those courses taken during the freshman year. In order to gain a handle on developing the needed methodology, a decision was made to work with a simplified version of the problem. Thus, analysis focused first on freshman year course data since this was the easiest situation conceptually.

The dimensionality of the freshman year course data was determined using principal factor analysis. However, some preliminary steps were necessary prior to implementing the principal factor analysis.

First, the number of courses to be analyzed had to be reasonably small in order for the analysis to be tractable. A minimum enrollment of 15 was used to select courses for inclusion in the analysis procedures. Using this criterion, 127 of the 664 courses, representing 87.5% of the 15452 freshman year course data records, were included in the analysis.

Second, a correlation matrix was computed on the grades students received in these 127 courses. This matrix of coefficients is incomplete. Because of structural limitations in the data, such as two courses being offered simultaneously, some coefficients cannot be computed since no student will have enrolled in both courses. Other coefficients are missing because from a practical point of view registering for certain pairs of courses (i.e., two different freshman English tracks) doesn't make sense. In order, then, to carry out the component analyses values were imputed for the missing correlation coefficients.

Two different imputation methods were used to complete the correlation matrix. The simpler method involved imputing a value of zero for all of the missing coefficients. The second imputation method required computing the corresponding row and column means for each of the missing coefficients. The overall mean of the correlation matrix was then subtracted from the sum of the row and column means. This value was calculated and then imputed for each of the missing coefficients. In other words, this imputed value is an average of the row and column effects for a given course.

This second imputation method leads to generally reasonable values because most of the non-missing correlation coefficients are moderate positive values in the range of .20 to .50, since these are correlations among grades. The imputed values then also fall into this range. Using either imputation method, the correlation matrix can be completed and analyzed. As it turns out, the two methods lead to quite similar results.

Principal factor analysis was carried out on two sets of freshman year courses. The first set consists of the 41 courses which had an enrollment of at least 100 students from this group. The second set consists of the 127 courses with a minimum enrollment of 15 students, including the 41 courses from the first set.

A correlation matrix was computed for each of the two sets of courses. Each coefficient is based on the correlation between grades for the corresponding sample of students enrolled in each pair of courses. One obvious problem with a correlation matrix computed in this manner is that each coefficient is based on a different self-selecting sample of students. It is not surprising that this kind of estimate of a correlation matrix may have negative eigenvalues.

The principal factor analyses are exploratory in nature and their results are meant to be used as a guide in partitioning courses in a sensible fashion. In part because of the unusual structure of the correlation matrix upon which the principal factor analyses are based, these results are meant to provide general guidelines rather than absolute rules for determining reasonably unidimensional sets of courses.

In all cases the method for extraction is principal factors followed by varimax rotation. For both the set of 41 courses and the set of 127 courses, the two factor solution is the most interpretable result. The most interpretable principal factor analysis solution is robust using different rotational methods and is robust for a wide range of obliqueness (including orthogonality) among the factors.

Using the results from the factor analyses, each unidimensional subset of courses is then scaled separately. The data matrix consists of rows representing students and columns representing courses. The cells of the matrix contains the grade of a particular student in a specific course. For all courses, there are only four possible grade values: 4 (which represents the grades of A+, A, A-), 3 (B+, B, B-), 2 (C+, C, C-), and 1 (D+, D, D-).

If a student did not receive a letter grade for a particular course, then his grade is 0 for that course. This value is not part of the ordered grade scale and is interpreted by the MSTEPS and ORDINAL programs as 'not presented'. 'Not presented' also includes those cases where a student does not receive a letter grade for that course (i.e., Incomplete, Pass, No Credit, etc.).

One ability estimate is computed for each student on each unidimensional set of courses. Students who have not taken at least one of

the courses in a set cannot have an estimate computed for that ability trait or component. A student's GPA for a particular set of courses is calculated as the average of the letter grades received in all of the courses which the student has enrolled. This differs from the actual GPA computed by the registrar's office where grades are weighted by the units of enrollment for each course.

The method used to obtain estimates on a set of courses assumes a prior distribution on each item parameter. The default distribution used by ORDINAL is the normal and assumes a mean of 1.0 for the slope parameter and a mean of 0.0 for the threshold parameter for each item. For both parameters, the standard deviation is assumed to be 2.0. These assumed values are reasonable and are the ones used in this study.

Due to the sparse nature of the data matrix of students by courses, MSTEPS is quite limited as to the size of the matrix that it can handle properly. Beyond a certain point, MSTEPS is unable to converge upon a solution and yield meaningful estimates. Depending on the specific courses included, the upper limit on the number of courses is around 20.

Thus, for any large set of courses, MSTEPS did not prove useful as a tool for scaling courses and persons. In fact, in analyzing courses beyond the freshman year, ORDINAL was the program of choice.

Validity studies were carried out using the three most commonly used preadmissions measures of HS GPA, SAT-V, and SAT-M as predictors and cumulative GPA, estimates of θ (the person ability estimate) from MSTEPS and ORDINAL as criteria in univariate multiple regression analysis. For this study, I have restricted analysis to only those predictors which have been traditionally used in most validity studies. There is a full range of other preadmissions measures that was not examined in this study.

Multiple regression analysis was carried out separately for each of the three subsets of courses. Within each subset of courses data from all four academic years was analyzed. In addition to ordinary least-squares (OLS) regression analysis, weighted least-squares (WLS) analysis is also done for each course subset.

WLS was used because estimates of performance are not equally precise for all students. In predicting GPA, the weight used for WLS is the square root of the number of courses in that area. In predicting the ORDINAL estimate, the weight used for WLS is the natural log of the inverse of the standard error. Transformations of raw weights were used because the large range of values for the number of courses and the standard errors would have given excessive leverage to certain observations. The transformed values produced a more reasonable set of weights.

FACTOR ANALYSIS RESULTS

The results from the principal factor analyses do not separate courses as cleanly as we would prefer, possibly because different

subsamples are used in computing each coefficient in the underlying correlation matrices. However, the results do give a good indication that it is reasonable to divide courses into one of two types: those courses that emphasize quantitative reasoning and deal with quantitative knowledge primarily, and those of a non-quantitative or verbal nature.

The solution used to guide the partitioning of courses is a two factor solution with all courses loading highly on only one of the factors. The two factors can be interpreted as a natural science and engineering factor and a social science and humanities factor. The courses that load highly on the second factor are courses in both the humanities and social sciences. However, results from the IRT programs indicate a lack of strong unidimensionality among these courses. Therefore, based on those findings, these courses were logically subdivided into two groups: humanities department courses and social sciences department courses.

Using a minimum enrollment of 15 students from this cohort, we have the following number of courses used to scale grades: 396 natural science and engineering courses, 184 humanities courses, and 198 social science courses, for a total of 778 courses across the four academic years. The allocation of departments is done on logical grounds and follows from the factor analysis results of freshman year courses. The departments of the three subsets of courses are listed below:

- 1) Natural Science and Engineering: Aeronautics and Astronautics, Applied Physics, Biochemistry, Biological Sciences, Chemistry, Civil Engineering, Computer Science, Electrical Engineering, Engineering-Economic Systems, General Engineering, Geology, Geophysics, Human Biology, Industrial Engineering, Mathematics, Mechanical Engineering, Operations Research, Petroleum Engineering, Mathematics, Physics, and Statistics;
- 2) Humanities: Art Studio, Art History, Asian Languages, Classics, Comparative Literature, Drama, English, French, German, History, Humanities, Italian, Linguistics, Music, Philosophy, Religion, Spanish, and Western Culture;
- 3) Social Science: Anthropology, Communications, Economics, Political Science, Psychology, and Sociology.

VALIDATION RESULTS

The scaling of individuals and courses is accomplished using either the MSTEPS or ORDINAL programs. MSTEPS was used only for the purposes of scaling freshman year courses. ORDINAL was used for scaling courses from the freshman year and from all four years.

The principal statistical procedure used in assessing the usefulness of the Universal Scale for Grades is multiple regression analysis. The independent variables are the three most commonly used preadmissions predictors of academic performance: SAT-V and SAT-M scores and high school

GPA. The criterion variables are:

1) The GPA based on the courses in a unidimensional subset that a student has taken. Unlike the actual GPA computed by the registrar's office, this is an unweighted average of courses, without regard for the number of academic units associated with each course. Weighting by units is useful for this study only if it leads to a more precise estimate of the overall quality of academic performance. In actuality, the weighted and unweighted versions of the GPA are very highly correlated. Note that course grades are not weighted by the IRT estimation programs, and so the unweighted GPA parallels the Universal Scale for Grades in that respect.

2) The estimate of a student's latent ability, θ , produced by MSTEPS or ORDINAL based on the grades earned in a unidimensional subset of courses.

For each pair of criteria derived from the same underlying set of courses, a regression analysis is conducted using the three predictors cited above. If adjusting grades for the difficulty of courses using IRT creates a measure of academic performance that is more reliable than the usual GPA, then we would expect that this new criterion would be more predictable from preadmissions information.

In fact, we find that this is the case here. Depending on the courses used to define academic performance, the increase in predictability of θ over the GPA is sometimes dramatic. This is documented in the table below:

Criterion Variables	Multiple R	R ²	Increase in R ² over GPA (OLS)
396 natural science and engineering courses (N=1462):			
GPA (OLS)	.53	.2774	
GPA (WLS)	.56	.3156	.0382
ORDINAL (OLS)	.61	.3729	.0955
ORDINAL (WLS)	.61	.3681	.0907
198 social science courses (N=1362):			
GPA (OLS)	.53	.2794	
GPA (WLS)	.55	.3066	.0272
ORDINAL (OLS)	.54	.2966	.0172
ORDINAL (WLS)	.56	.3136	.0342
184 humanities courses (N=1474):			
GPA (OLS)	.52	.2672	
GPA (WLS)	.51	.2638	-.0034
ORDINAL (OLS)	.51	.2617	-.0055
ORDINAL (WLS)	.52	.2687	.0015

At the present time, no formal statistical test exists for testing the difference between two sample multiple correlation coefficients based

on different linear combinations of predictors.

As shown above, the degree of improvement in using the Universal Scale for Grades over GPA depends on the courses used to measure performance. There are both substantive and statistical reasons that account for this. The improvement is largest for the natural science and engineering courses, sizable for the social science courses, and marginal for the humanities courses.

The large observed improvement of the ORDINAL estimates over GPA based on the natural science and engineering courses may be due to the fact that the underlying ability or trait that determines performance in this area is better defined and thus easier to measure than for the social science or humanities courses. Additionally, there are statistical reasons that may contribute to this result.

First, the distribution of grades in the natural science and engineering courses is more variable. Additionally, the problem of grade inflation where all or almost all of the students in a course receive A's is especially true for many of the humanities courses. Second, the precision of the ORDINAL estimates is best for the case of the 396 natural science and engineering courses. There is a larger pool of courses to utilize, one that is at least twice as large as the other subsets of courses. From this larger pool of possible courses, each student takes more courses than in the other two areas.

The difference in the precision of ORDINAL estimates can be summarized by the average standard error of estimate for the three subsets of courses: .27 for the natural science and engineering courses, .43 for the social science courses, and .40 for the humanities courses.

Estimation of performance for the humanities courses is the most difficult of the three cognate areas. The defining trait in humanities courses may be 'fuzzier' than in the other cases. Prediction of performance for social science courses falls between the natural science and engineering and humanities courses. This seems to be consistent with the substantive and statistical comparisons among the three subsets of courses.

Alternatively, estimates of the variance components for each of the three subsets of courses were computed. In this analysis, courses and persons are each assumed to be randomly selected from an infinite population. The goal of this analysis is to estimate the variance of each of these population. The estimates of the variances obtained are as follows:

Source	184 Humanities Courses	198 Social Science Courses	396 Nat Sci Courses
Courses	.035	.100	.070
Students	.100	.240	.230
Error	.250	.280	.345
Total	.385	.620	.645

Using these figures, the reliability for a single course can be computed. The reliability for a single course is .26 for the humanities courses, .39 for the social science courses, and .36 for the natural science and engineering courses. Thus, the lower reliability of the humanities courses may contribute to the lower predictability of the ORDINAL estimates based on those courses.

It is also important to examine the interrelations among the different criteria and predictors. Table 1 of the Appendix shows the matrix of correlation coefficients among the three ORDINAL estimates, the three GPA's, and SAT-V, SAT-M, and high school GPA.

In general, each of the preadmissions predictors correlates better with the ORDINAL estimate than with the GPA. This is true for each comparison of the three pairs of criteria. The best single predictor of performance in the natural science and engineering courses is SAT-M, while the best predictor of performance in the humanities courses is SAT-V.

Note that high school GPA is never the best single predictor, and is always exceeded by one, usually both, of the test scores. The incremental gain of SAT scores over high school grades in predicting each of the criterion measures is shown in the table below:

Predictor Variables	Correlation with:	
	GPA (OLS)	ORDINAL (OLS)
396 natural science and engineering courses (N=1462):		
HS GPA only	.33	.37
SAT-V, SAT-M, and HS GPA	.53	.61
198 social science courses (N=1362):		
HS GPA only	.36	.38
SAT-V, SAT-M, and HS GPA	.53	.54
184 humanities courses (N=1474):		
HS GPA only	.33	.30
SAT-V, SAT-M, and HS GPA	.52	.51

These results show that SAT scores contribute substantially to the prediction of long-term academic performance above and beyond the predictive validity of high school grades. The degree of improvement from using the SAT is somewhat better when the criterion is the ORDINAL estimate. For this group of Stanford students, prediction of their academic performance at Stanford is greatly enhanced by the use of SAT scores.

Part of the explanation of why test scores are substantially better than high school grades as predictors lies in the distributions of the

variables for this cohort. The distribution of high school GPA is highly positively skewed for this group of students, which is not surprising given the selectivity of the university. Many students have perfect high school GPA's of 4.0. Thus, high school GPA when used as a predictor of future performance suffers from a restriction in range of the variable. SAT scores, while generally also high, do not suffer as much from the problem of restriction in range as the high school GPA.

Restriction in range accounts only partly for the poorer quality of high school grades as a predictor of performance in a unidimensional subset of courses. High school GPA suffers from the same problem that the college GPA does in that it is also a multidimensional measure of performance distilled into a single number. On the other hand, SAT-V and SAT-M scores, taken separately, are in some sense more pure measures of performance in a unidimensional area. It is not surprising, then, to find that the best predictor of performance in the natural science and engineering courses is the SAT-M score while the best predictor of performance in the humanities courses is the SAT-V score. Performance in social science courses is about equally predictable from SAT-V, SAT-M, and high school GPA.

SUMMARY AND DISCUSSION

This dissertation has sought to address an important and long-standing problem in educational measurement: What is the predictive validity of traditional preadmissions measures in determining overall academic performance in college? The study provides new insight into this question via a new application of an existing statistical methodology, the technology of Item Response Theory (IRT).

This is the first study which uses IRT methods to adjust the cumulative GPA in order to better address the issue of the predictive validity of college admissions measures. This study is an inquiry into the feasibility of this approach and should be viewed as a prototype for further research.

IRT is used to estimate the latent abilities of a student based on the relative difficulties of the courses he or she has enrolled and the grades received in those courses. In this way, a new measure of performance is developed, one that may be thought of as an adjusted GPA. As expected, this approach defines a measure with greater reliability than the cumulative GPA normally computed. This new measure of academic performance, the Universal Scale for Grades, is more reliable in the traditional psychometric sense. This greater reliability is manifested in greater predictability from the three most commonly used preadmissions variables: SAT-V, SAT-M, and high school GPA.

Prior attempts to adjust grades for differences in grading standards using other methods have yielded marginal improvements. This study is a break from the traditional approaches to grade adjustments and represents a substantial improvement over the results obtained in the past. The use

of IRT methods in this context shows significant potential in increasing our understanding of predictive validity in college admissions.

This study used data from two sources at Stanford University: the undergraduate admissions office and the registrar's office. The cohort under study is the entering freshman class of 1982. The data base provides a wealth of information on each student's demography, preadmissions variables, and Stanford course history.

This study utilized computer programs which implement the the two unidimensional IRT models, the Partial Credit Model and the Graded Response Model. The corresponding programs are called MSTEPS and ORDINAL, respectively. Principal factor analysis is used to partition a larger set of courses into approximately unidimensional subsets. Each subset of courses is then scaled independently using the two programs. These programs produce item and person parameter estimates. SAS is used for file creation and management and for statistical analysis.

The results indicate that use of IRT methods to create more reliable measures of academic performance shows promise. Each outcome measure is based on a unidimensional set of courses, and collectively, are called the Universal Scale for Grades. The Universal Scale for Grades has better reliability than the traditional GPA because the estimates of students' performance is based on the relative scaling of an approximately unidimensional set of courses.

The improvement in predictability, that is the increase in R^2 , of the Universal Scale for Grades over the cumulative GPA ranges from .0055 to .0955 using the three predictors of SAT-V, SAT-M, and HS GPA. The degree of improvement depends on the particular set of courses under consideration. There are a number of alternative psychometric models to the ones used in this study, and the utility of these alternatives are yet to be explored.

The sparse nature of the data matrices used in this study dictates the use of certain options in the estimation algorithm. However, there are alternate ways of viewing the data that could enable us to explore these other models and options.

This study shows that significant increases in the predictability of academic performance are possible. By improving the reliability of a criterion measure, the predictive effectiveness of certain preadmissions measures will increase. This has been shown to be the case. Alternatively, traditional validity studies may be interpreted as understating validity coefficients due to correctable defects in the GPA as a criterion.

The degree of improvement appears to be related to the dimensionality of the courses under consideration. Subsets of courses that are closest to unidimensionality benefit most from the use of IRT methods. In this study, it is the natural science and engineering courses that yield the greatest increase in predictability. On the other hand, the humanities courses show the smallest incremental gain from the new

methodology. Additionally, the larger the pool of courses used to scale grades, the greater the improvement from using IRT methods. Some preliminary findings using an expanded set of social science courses indicate that this is so.

There are limitations to the inferences we can make based on the findings of this research. Some of these limitations are stated explicitly below:

1) IRT methods are known to have applications in many testing situations. We do not know how well these methods map onto the type of problem we are addressing here. In particular, the effects of person by course interactions on the results is not known, and cannot be easily determined. These interactions introduce a form of departure from unidimensionality and may serve to systematically bias the results. At this time, we do not know how robust IRT methods are to departures from unidimensionality when used in this context.

Also, the IRT models used in this study assume conditional independence of items. This assumption manifests itself in a number of ways. For example, it is assumed that an examinee does not gain information about later items from prior items. We assume that significant learning does not occur during the test administration. Clearly, this differs from what we know happens as students progress through college. It is desirable for students to utilize their knowledge from past courses in future courses. The assumption that no learning occurs becomes increasingly less tenable as we examine data from multiple academic years.

2) We do not know the impact of self-selection by students into courses on the validity of using IRT methods. Self-selection by students introduces another source of bias into the study's findings. Some students are better than others at identifying those courses in which they will do well. It is possible that unidimensional models need to take into account this additional trait.

3) The IRT models used in this study make assumptions about the data. Many of the results in this study are based on the simplest form of the GRM where courses are assumed to share a common set of categorical parameters. In other words, we assume that the relative distance among category boundaries on the θ scale is constant for a set of courses, and that courses differ only on location on this scale. This is a restrictive assumption which limits the degree of improvement possible using the methods in this study. In the future, if we can obtain reliable estimates using the unconstrained version of the GRM, then the reliability of the Universal Scale for Grades will increase. This would further improve the predictability of the Universal Scale for Grades.

4) This study examines one cohort at one university. How well do these same methods work on different kinds of data? Cross-validation and replication is one of the cornerstones of science. Does the validity of the methods hold for different cohorts of undergraduates at Stanford? How well do these methods work with data from other institutions? After all,

Stanford is not typical of most American universities. In addition, what are the limits, in terms of the number of students and courses and the size of courses, that can be addressed using these methods?

The findings from this dissertation have implications in several areas in educational research. In the area of educational measurement, it appears that IRT models have wider applicability than previously believed. This study shows that the use of IRT models for equating grades from different courses onto a common scale is a fruitful new application.

In the area of institutional research, this study shows that developing a more reliable criterion of academic performance leads to significant improvements in the predictive power of common preadmissions measures. For a single cohort of students at one university, I show that the three predictors of SAT-V, SAT-M, and HS GPA are significantly better at forecasting longer-term measures of academic performance than has been shown previously.

Thus, it appears that the power of the traditional preadmissions measures to predicting performance through all four years of college is better than we have thought to be true. This has implications for undergraduate admissions offices at many selective institutions. If academic performance is a criterion for selecting applicants for admissions, performance measured not just by first year grades but by four years of course-taking, then this study shows that, for this cohort, the combination of SAT-V, SAT-M, and HS GPA works very well.

Critics of standardized tests such as the SAT may need to reconsider their contention that the SAT are valid only for predicting first-year grades and that the tests have no proven validity in predicting longer-term measures of academic performance. The evidence from this study indicates that SAT scores predict four-year cumulative performance well if the criterion measure has improved reliability.

This is not to imply that predicted grades should be the only criterion by which applicants are selected. On the contrary, there should be a number of other dimensions on which applicants should be evaluated. But if overall academic performance is an espoused criterion for admissions, then we can be reasonably confident that the traditional preadmissions measures predict this criterion to a remarkable degree.

I conclude by reiterating that this study is only the first step in using IRT methods to obtain a better understanding of predictive validity in undergraduate admissions. This study demonstrates that this approach appears promising, and that traditional preadmissions measures perform better than previously expected in predicting a longer-term measure of academic success.

TABLE 1
CORRELATIONS AMONG IRT ESTIMATES,
GRADE POINT AVERAGES AND PREADMISSIONS MEASURES

Key to variables:

- 1 = ORDINAL estimate based on 396 natural science and engineering courses
- 2 = ORDINAL estimate based on 198 social science courses
- 3 = ORDINAL estimate based on 184 humanities courses
- 4 = GPA based on 396 natural science and engineering courses
- 5 = GPA based on 198 social science courses
- 6 = GPA based on 184 humanities courses
- 7 = SAT Verbal score
- 8 = SAT Mathematics score
- 9 = high school GPA

	1	2	3	4	5	6	7	8	9
1									
2	.65								
3	.53	.58							
4	.90	.65	.55						
5	.63	.88	.56	.63					
6	.52	.57	.87	.55	.55				
7	.38	.41	.46	.36	.42	.45			
8	.56	.43	.30	.46	.40	.32	.47		
9	.37	.38	.30	.33	.36	.33	.24	.30	

REFERENCES

- Bloom, B. S., and Peters, F. R. Academic prediction scales. New York: The Free Press of Glencoe, 1961.
- Boldt, R. F. Efficacy of undergraduate grade adjustment for improving the prediction of law school grades (Report No. LSAC 76-4). In Law School Admissions Council, Reports of LSAC Sponsored Research: Vol. III, 1975-1977. Princeton, NJ: Law School Admissions Council, 1976.
- Linn, R. L. Grade adjustments for prediction of academic performance: A review. *Journal of Educational Measurement*, 1966, 3, 313-329.
- Masters, G. N. A Rasch model for partial credit scoring. *Psychometrika*, 1982, 47, 149-174.
- Muraki, E. ORDINAL (Computer program). Chicago: Scientific Software, 1988.
- Rosenbaum, P. R., and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 1983, 70, 41-55.
- Samejima, F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, Monograph Supplement No. 17, 1969.
- Tucker, L. R. Formal models for a central prediction system. *Psychometric Monograph No. 10*, Richmond, VA: William Byrd Press, 1963.
- Wilson, K. M. A review of research on the prediction of academic performance after the freshman year. College Board Report No. 83-2 and Educational Testing Service Research Report No. 83-11. New York: College Entrance Examination Board, 1983.
- Wright, B. D., Congdon, R. T., and Rossner, M. A user's guide to MSTEPS: A Rasch program for ordered response categories (Computer program and user's guide, Version 2.1). Chicago: Scientific Software, 1987.
- Young, J. W. Developing a universal scale for grades: Investigating predictive validity in college admissions. Doctoral dissertation, Stanford University, 1989.